## The Data Management Issue for WCRP

Bryant J. McAvaney (JSC/CLIVAR WGCM)

### Introduction

The complexity of the climate system is reflected in the huge volume of data that is collected and stored at a large number of different locations under the WCRP umbrella. In order to utilise this data in an efficient and timely manner there is a great need for a data infrastructure that spans the many countries, agencies and institutions.

Within WCRP climate data is held in a huge range of technologies from huge relational databases to files on an individual scientist's PC. The data is often relevant to a wide range of scientific disciplines despite often having been collected on behalf of quite narrow specialist domains. As Earth Systems Science continues to evolve the requirement for inter-disciplinary data access will increase. Individual scientists and individual research groups will increasingly need better means of locating data and having it delivered than exist at the moment. A major problem is that data is becoming more and more distributed but there is a lack of interoperability between different data archives so that data access is becoming increasingly frustrating. What is needed is some coordination of data structure and data management that extends far beyond traditional organisational and national boundaries.

All scientific data users go though a number of steps outlined in Figure 1 [1] in order to utilise digital data. While not all steps may be needed in any one application and the order may vary each of the steps can itself consist of a complex set of operations especially from discovery to extraction and processing and display. More often than not more than one data set is used requiring different tools at each step. Perhaps the most difficult step of all is getting started.  Data discovery often relies on word-of-mouth to find out about new or newly reprocessed data.  Even when sources of data are known, the user must usually find and locate specific files or databases of interest and then go through the process of learning about data formats and or appropriate data base schema before using the data for scientific purposes. Much data is simply not used outside of the local generation site because of the huge overheads of locating and handling it. Too much trained scientist time is being spent "reinventing wheels".

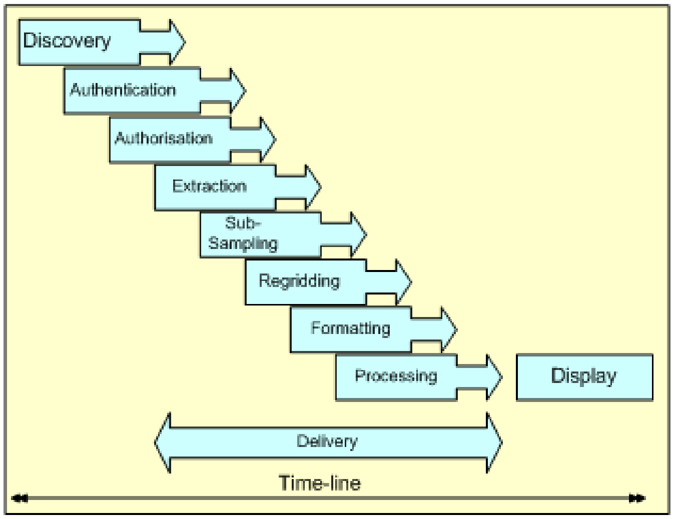

**Figure 1 – The data discovery and usage chain**

## A vision of Data Usability.

At the present time the situation with regard to data is similar in many ways to the state of access to documents before the advent of the Web and its search engines. In pre-Web days, access to documents involved locating them, transferring the needed files, converting from one format to another and transforming document excerpts into common forms suitable for reading or merging into papers, presentations or reports.

In the not-too-distant future, access to data for analysis and visualization could be almost as simple as access to documents is today through a Web browser interface. Just as our networks and web browsers now interoperate to support location transparency, so our programs should be able to access, analyse, visualize, and integrate data from either local or remote sources. In the same manner that HTML has become the lingua franca of the web that enables anyone to publish documents, a standard metadata architecture could provide the means by which our important datasets could be easily published for access by local and remote applications, catalogued by search-engine services, and found by web browsers and other applications.

The existence of the Web and browsers provided benefits that encouraged the use of HTML and web servers which multiplied as use of the web became more widespread. Similarly a "data Web" constructed from existing component protocols and formats for remote data and metadata access could be combined into a framework that would increase benefits as usage increased. The data web will only become a reality when it is as easy to publish data with metadata and data services that make it as useful as it is to publish documents on the Web. WCRP could play a role in catalyzing such a vision.

A continuing problem is of course the resources needed to realise the vision. This is particularly difficult in the current context since funding agencies tend not to support activities that provide highly general solutions that are reusable in other contexts and the construction of generally useful frameworks is a very challenging task. Similary data providers usually have no incentive to provide the extra metadata information and organisation of their data that would make it useful in unanticipated contexts. WCRP could alert national funding agencies to the importance of general solutions.

## Strategy

There are a number of initiatives being taken that address many of the issues outline above [2] [3] [5] [6]. There are also a number of significant efforts taken by organisations to provide portals, catalogues and gateways to environmental and atmospheric data resources. Links to some of these (mostly USA) can be found from http://www.unidata.ucar.edu/staff/russ/dmwg/portals.html

A very important consideration is how WCRP should interact with these various national and organisational efforts. Two extreme approaches would be to 1) promote the centralised development the necessary infrastructure with a unified collection of data assets and services, or 2) promote the continuation of separate and nearly autonomous activities, each with their own tailored data standards and policies for sharing data.

The first extreme would standardize access to the diverse data assets within the WCRP programmes by promoting a single WCRP wide format and virtual data portal, independent of discipline boundaries or user requirements, making it possible to monitor usage, standardize protocols, and encourage metadata standards and policies across all WCRP programmes. Such an effort would develop an organized data collection from numerous existing collections of observational data, model outputs, field experiments, and derived, value-added, data. It would require not only organizing existing data assets, but also constraining new data assets and services to fit within a WCRP integrated data management standard framework that emphasized interoperability.

At the other extreme, which resembles the status quo, each WCRP programme responsible for data assets would make these available individually to other appropriate organizations, disciplines, and users independently of other efforts within WCRP. Rather than choosing global standards for the representation of metadata or interfaces for data access, use of standards most appropriate for each specific data resource or service would be encouraged. For example, archived climate data would be made available through servers and application programming interfaces most familiar to climate researchers, whereas real-time data from a field experiment would be provided in a form most useful to the investigators involved with the experiment. The fact that WCRP programmes made both data sets available would not influence the technical decisions that determine how those data sets were organized.

The first extreme represents a level of centralization and standardization that is neither practically achievable nor desirable. Constructing or selecting suitable standards to encompass all the data within WCRP is an almost impossible task; by the time all existing WCRP data collections were moulded to make them conform to selected standards, those standards might be obsolete and the cost in terms of personnel time would very large. The provenance of data is an important data attribute, but it should not be the primary attribute determining the organization or representation of the data, especially when the effort to do so stifles innovation.

The second extreme, complete data autonomy is also undesirable for several reasons. It requires every group and project responsible for data within WCRP programmes to design and implement their own means for making that data available to others, including awareness of the best practices for metadata representation for discovery and use, knowledge of how to make the data useful to a larger set of current and future uses than the specific project that generates the data, and resources for providing all the data services that are needed for efficient access to the data. Data autonomy and lack of resources lead to lack of awareness of beneficial connections among the data collections, and no way to readily determine how to access data from other groups in the organization. The scientific scope in research projects is broadening, not becoming more focused. To remain in concert with the scientific needs WCRP must further integrate our available data resources.

Instead of either of these extremes, encouraging the loose combination of legacy systems while encouraging the development of new ways to support data access to WCRP data assets would permit national agencies to continue to work on the cutting edge of distributed data systems. To achieve this interoperability, significant effort will be required from programmers and scientists throughout WCRP programmes.

### *Specific Requirements*

WCRP should foster interoperability by encouraging the integration of existing systems that have already ‘proven their worth’ and encourage national and organisational decisions about systems that have evolved and successfully occupy a ‘data niche’. Imposing a “top-down” set of standards would be counter productive.

#### *Data Discovery and Data Extraction*

Supporting data discovery is a complex task and depends on the existence of metadata. In its broadest sense metadata are simply *“a range of structured ancillary information about data”* which describes the attributes of an information resource. One simple example is the description of products in a hardware catalogue. For a climate researcher an example is the description of the observing practice at an observational site or the header files describing the gridded output from a reanalysis.

Metadata can be conceptually classed into two general types, discovery and use. Discovery metadata addresses the information necessary to identify a data collection and determine whether it is available and appropriate for the intended application. Use metadata provides the technical information necessary to actually use the data in the collection. Of the two types, use metadata are more mature due to the creators and users of geodata converging in the last decade to a modest number of data storage formats containing reasonably well defined data descriptions. Discovery metadata has only recently become an issue as operational and science centres have begun to move from static, in-house, data archives to more dynamic, online, data services.

Efficient exploitation of massive data sets requires cataloguing and documentation through the use of metadata, i.e., data describing the primary data objects themselves. In addition, verifiability of simulation-based research requires systematic collection and maintenance of metadata that document the design and execution of a simulation or collection of simulations. Locating science information within the massive data archives is currently difficult and requires considerable intimate knowledge of the organization and structure of the archive. To facilitate discovery, metadata must be standardized and organized into databases that support a variety of query types. Different classes of queries require different types of metadata to identify information such as what data are available, the nature of the data, how they were generated, and where they are located. Current metadata conventions used in the community (COARDS, CSM, GDT, SOHO-FITS, CEDAR, etc.) address primarily the description of the contents of individual files. These conventions need to be extended to encapsulate information about data collections and their derivation history. For example, environmental simulation systems are often composed as distributed applications; each component can represent a physical subsystem, such as the atmosphere, the ocean, or a level in a grid hierarchy. Each component may be responsible for its own output processing. Metadata must identify the relationships between the components to allow reconstruction of the overall simulation configuration. Similarly, data will often pass through many post-processing steps after the completion of the simulation. Each of these steps needs to be documented in the metadata. Identifying the appropriate common semantics and granularity of discovery metadata, upgrading legacy use metadata for online applications and ensuring that metadata are retained in a dynamic environment are all topics to which WCRP programmes will need to be addressed if successful implementation of location independent data services across distributed data centres is to be achieved.

The production of metadata by a data provider (which requires commitment over and above provision of the data itself) must be rewarded in some way. Organisations and individuals will need to be gain better portals into their own data as well as that of others. WCRP can assist by promoting the role of data providers as part of an overall data services strategy. It is assumed that the aim is the provision of high quality data that is visible and available to the research community at large. Data providers need to be encouraged to take advantage of existing mechanisms and technologies to make their data more accessible and hence serve the wider community. Encouraging data providers to develop useful metadata using widely-used conventions will mean that valuable information will be provided to researchers and this information can be more easily found using advanced data search tools. In many cases data providers should be encouraged to provide their data using a client/server data access model instead of a file based system; this would give the advantage that it facilitates access to datasets that is physically separate from the actual location of the data.

### Training

WCRP should encourage the dissemination of information regarding the need for interoperability of data provision and the relevant software engineering that is required.

### Recommendations

1. WCRP should coordinate activities within its programmes so as to facilitate a vision of data usability that is suited to the Web.

2 WCRP should encourage national funding agencies to support the development of general data management tools that benefit the entire scientific community.

3 Conduct a WCRP Workshop on Data Management that brings together policy advisors and software engineers so that details of a WCRP "vision" can be explored.

### *Acknowledgements*

This paper has benefited from many informal discussions with a wide range of people who have pointed the author to many different sources. The assistance and material provided by Glenn Rutledge, Jonathon Gregory and Bryan Lawrence was especially important during the development of this text. This paper rests on the ideas that are at the core of the NOMADS initiative in the USA, the NERC DataGrid in the UK, the Earth System Grid Project in USA and the PRISM initiative in Europe.

### References

[1] The NERC DataGrid (http://www.w3c.rl.ac.uk/Euroweb/poster/118?EuroWeb2002PosterAbstract2.htm/)

[2] The NERC Metadata Gateway (http://www.nmp.rl.ac.uk/ )

[3] The CLRC Data Portal project (http://esc.dl.ac.uk:9000/index.html )

[4] Open Grid Services Architecture Database Access and Integration (http://imbriel.dcs.gla.ac.uk/NeSC/general/projects/OGSA_DAI/ )

[5] The Earth System Grid II project (http://www.eathsystemgrid.org/ )

[6] NOMADS: The NOAA Operational Model Archive and Distribution System (http://www.ncdc.noaa.gov/oa/model/model-resources.html)

[7] DODS/OPeNDAP: Open Source Project for a Network Data Access Protocol (http://www.unidata.ucar.edu/packages/dods/home/swODC/ )

[8] Unidata (http://www.unidata.ucar.edu/staff/russ/dmwg/portals.html )